Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion

Emiel Krahmer, Erwin Marsi, Paul van Pelt
Tilburg University
The Netherlands

Plan

1. Introduction: sentence fusion

2. Q-driven vs. Generic sentence fusion
   - Experiment 1: Data-collection
   - Experiment 2: Evaluation

3. Summary and outlook

Sentence fusion

- Sentence fusion: given two related sentences, produce a single sentence containing the shared information (Barzilay et al. 1999, Barzilay & McKeown 2005)
- Text-to-text generation
- Motivation: Beneficial for multi-document summarization. Less redundancy, more informative summaries (Barzilay & McKeown 2005)
- QA applications: fuse alternative answers to obtain a more complete answer

Example: Generic fusion

- Answer 1: Posttraumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.
- Answer 2: Posttraumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.
- Fusion: Posttraumatic stress disorder (PTSD) is a psychological disorder.

Complication

- Daume III & Marcu (2004): “Generic sentence fusion is an ill-defined summarization task.”
- When participants are asked to fuse two consecutive sentences from a document, their results are widely different.
- If even human participants don’t agree, evaluating sentence fusion is tricky...

Our solution/hypothesis

- Query-based fusion: Fusing two answers guided by a question
- Hypothesis: Query-based fusion gives a higher agreement on the task

Example: Query-based fusion

- Question: What is PTSD?
- Answer 1: Posttraumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.
- Answer 2: Posttraumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.
- Q-based fusion: PTSD stands for posttraumatic stress disorder and is a psychological disorder.

Fusion types

- Marsi & Krahmer (2005): There is more than one way to fuse two sentences.
- Intersection Fusion: only information shared by both sentences
- Union Fusion: all information from both sentences (but without redundancy)
- Which type of fusion is best for a particular application is an open question...

Example: Intersection vs. union fusion

- Answer 1: Posttraumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.
- Answer 2: Posttraumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.
- Intersection Fusion: Posttraumatic stress disorder (PTSD) is a psychological disorder.
- Union Fusion: PTSD (posttraumatic stress disorder) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Perspectives

- Generation perspective:
  - Is Q-based fusion a better defined task?
  - Will people agree more on union than on intersection fusions?
  - Is the effect of the preceding question the same for both unions and intersection fusions?
- User perspective:
  - Do users prefer concise (intersection) or complete (union) answers?
  - And does it matter whether they were generic of Q-based?
- Next: two evaluation experiments which address these questions...

Experiment 1: Data collection

- Materials:
  - Used QA benchmark set (100 questions, medical domain).
  - Correct answers were manually retrieved from the text corpus.
  - Selected 25 questions with multiple answers, with at least some shared information among answers

- Task: first perform generic fusion; next Q-based fusion

- Mixed between-within participants design. Two between conditions: Intersection and Union. Within each condition, both Generic and Question-based.

Experiment 1: Data collection (cont'd)

- Participants: 44 participants (24 men), average age 30.1 years. Randomly assigned to conditions.
- Method: web-based script.

Results (1)

- Descriptive statistics

Fusion Type	Length M (SD)	# Ident.
Q-based Intersection	8.1 (2.5)*	189*
Generic Intersection	15.6 (2.9)	73
Q-based Union	19.2 (4.7)*	134^
Generic Union	31.2 (7.8)	109

** p <. 001, ^ n.s.*

Results (2)

- (Normalized) ROUGE scores

	Generic Intersection	Q-based Intersection	Generic Union	Q-based Union
Rouge-1	.036	.068	.035	.041
Rouge-SU4	.014	.038	.018	.020
Rouge-SU9	.014	.040	.016	.020

In sum: Generation perspective

- Q-based fusions are shorter display less variation in length, yield more identical results, and have higher ROUGE scores.
- So: Q-based fusion is indeed a better defined task.
- But: does it matter?

Experiment 2: Evaluation

- Materials:
  - Selected 20 questions for which multiple (different) answers were obtained in Experiment I.
  - Per question, 4 representative answers were selected from the data collection, one for each category: Q-based Intersection, Q-based Union, Generic Intersection, Generic Fusion.

- Within participants design. For each of the 20 questions, participants have to rank the four answer (forced choice paradigm)

- Participants: 38 participants (17 men), average age 39.4 years.

- Method: simulated medical QA system

Results

- Average rank

1	Q-based Union	1.888*
2	Q-based Intersection	2.471*
3	Generic Intersection	2.709*
3	Generic Union	2.932

** p <. 001*

In sum: user perspective

- Q-based answer fusions are systematically preferred over generic ones.
- Comprehensive (union) answers are preferred over concise (intersection) ones

Summary

- Is Q-based fusion a better defined task?
  Yes. Q-based fusions are shorter, less varied, yield more identical solutions and have higher (normalized) Rouge scores than their generic counterparts.

- Which type of fusions do users prefer in a QA context?
  Q-based Union >> Q-based Intersections >> Generic Fusions

- Future work:
  - Follow-up experiments looking at the influence of question wording and at different domains
  - Working on extended fusion algorithm, based on Marsi & Krahmer (2005)